iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** I **Month of publication:** January 2021

DOI: https://doi.org/10.22214/ijraset.2021.32690

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue I Jan 2021- Available at www.ijraset.com

COVID-19 Outbreak Prediction Analysis using Machine Learning

Sagar Bala
Department of Computer Science, SRM Institute of Science and Technology

Abstract: ***The global spread of the novel coronavirus SARS-CoV-2 has resulted in the outbreak of a respiratory illness known as COVID-19. The coronavirus COVID-19 pandemic is the greatest global health crisis faced since World War II. Countries are trying to slow the rate of the outbreak by testing/treating their patients and putting total country lockdown i.e by quarantining citizens, limiting travel, closing all social gathering places likes: Temples, Colleges/Schools, Offices, etc. This pandemic has caused social, economical, political instability in every country.COVID-19 outbreak prediction analysis can help to predict this pandemic, recognize the spread cycle pattern which can help to take appropriate measures to lower the spread rate and use of the medical and economic resources effectively to areas of greatest need by public Health Officials, Ministry of Health, WHO. The proposed model called "COVID-19 Outbreak Prediction analysis using Machine learning" analyses how popular supervised learning regression models like Linear Regression, Support Vector Machine Regressor, Random Forest Regressor, and XGBoost Regressor can predict COVID-19. The proposed model aims to predict the Outbreak in advance for any particular country and compares the performance. It is observed that the performance of XGBoost Regression gives the best performance and almost comparable to Random forest Regression followed by the Support Vector Regression and Linear Regression model. However, accuracy can be improved using more accurate data in the future.***

Keywords: COVID-19, Machine Learning, Outbreak, Linear Regression, Support Vector Machine Regressor, Random Forest Regressor, XGBoost Regressor, Ministry of Health, WHO.

I. INTRODUCTION

This section discusses the recent researches done on COVID-19 that have used Machine learning algorithms to predict various factors of Covid19.

Centralized data Collection of COVID-19 patients is important for prediction and diagnosis for COVID-19, Ahmad Alimadadi,2020[1]. Developed or validated model for multilinear COVID-19 prediction and concluded that the model cannot be used in current medical practice for prediction due to its high risk of bias, poor reported performance and it should be considered candidate predictors for a new model. It needs a more rigorous and validating model for prediction, Wynants,2020[2]. How control measures impacted the containment of the epidemic in Wuhan, China using the SEIR model. The model was effective in predicting the epidemic peaks and sizes of COVID-19, YANG,2020[3].Prediction of the criticality of the patients suffering from COVID-19 using machine learning prognostic model and clinical data of Wuhan, Yan,2020[4].To develop a COVID-19 detection fully automatic framework using chest CT and Deep learning method, Li,2020[5]. COVID-net using deep convolution neural network for detection of COVID-19 from chest X-ray, Wang,2020[6].SEIR model and Regression model have been used for predictions based on the period of 30th January 2020 to 30th March 2020, Pandey, 2020 [7] .comparative analysis of machine learning and soft computing models to predict the COVID-19 outbreak using Multi-layered perceptron(MLP) and adaptive network-based fuzzy inference system(ANFIS). The results of two ML models (MLP and ANFIS) reported a high generalization ability for long-term prediction, Ardabili, 2020[8].Use of convolution neural network models for the detection of coronavirus pneumonia infected patients using chest X-ray radiographs, Narin,2020 [9].Implementation of a universal model for prediction using Gaussian function and chi-square distribution function named H-gaussian model, Wang,2020[10].The predictive machine learning model for COVID-19 using cross-validation on the routine blood tests of 5,333 patients with various bacterial and viral infections, kukar, 2020 [11].Developed a machine learning model based risk prioritization to predict ICU transfer within 24hrs between 26 February 2020 till 18 April 2020.Cheng,2020[12].

The algorithmic models used in this review are the most powerful and effective models whose real-world applications are in every sector such as Health care, Technology, Transport, Communication, Agriculture, Commerce, etc.Linear Regression model is used in dynamic best selling prediction in E-commerce for enhancing product search, Long,2012[13]. Author age prediction using Text, Nguyen,2011[14].Support Vector Regressor model is used in prediction for the converter gas tank level, Han,2012[15].

Earthquake prediction model, Asim KM,2018[16].Random forest regressor is used in the outbreak prediction of avian influenza H5N1, Kane,2014[17].MRI image synthesis, Jog,2017[18].XGBoost regressor model can be used for Chronic kidney disease diagnosis, Ogunleye,2019[19].Social media popularity prediction,Li,2017[20].

To curb high-risk bias and poorly reported models. Therefore, predictions are made using the algorithmic models, and their accuracy and error rates are compared to determine the most efficient model for Covid-19 outbreak prediction.

II. PROPOSED WORK

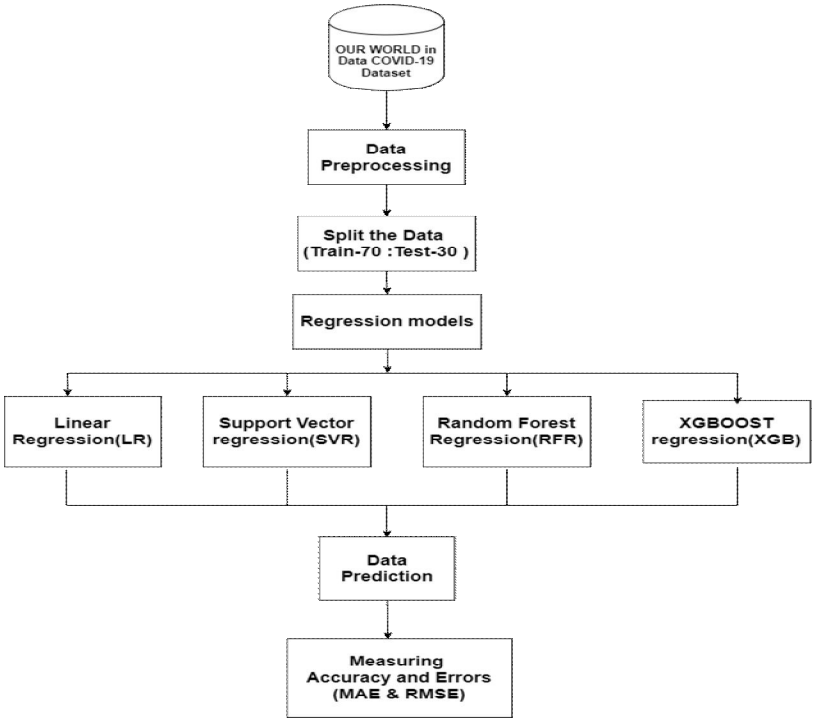


Fig. 1. COVID19 prediction model Architecture

All Our world in Data COVID-19(OWID) dataset on confirmed cases and confirmed deaths are being updated daily and are published by the European CDC under the university of OXFORD, the best available global dataset on the pandemic. They license all charts under Creative Commons by and they can be embedded in any sit. The total number of cases registered in the dataset is 69, 60, 259. The proposed model and simulation of data is done in Google Colaboratory notebook using python 3.7 languages. Machine learning has three aspects but in this study only supervised learning is used under the regression model. The data has been preprocessed and split into a 70: 30 ratio for training and testing respectively. For Prediction, the date is taken in the independent axis and total cases in the dependent axis. The prediction is measured using MAE and RMSE performance metrics and Score.

Since the proposed model Fig.1 makes use of Machine learning algorithms namely, Linear regression, Support vector regression, Random forest regression, and XGBoost regression. The basic understanding of these algorithms is discussed below.

A. Linear Regression (LR)

Linear regression is used to predict the values of dependent variables using the independent variables provided there exists a linear relationship between them. Let X be the independent value and "Y " be the dependent value, The Y value is predicted using the Eq.1.

$Y = a_0 + a_1 X$

Eq.1. LR equation

Where Y - variable represents the total number of cases and X variable represents the date,a0 denoted the Y-intercept and a1 indicates the slope. The linear regression model is built by learning the values of a0 and a1 from the training data set and predict the output for any given date.

B. Support Vector Machines (SVR)

Support Vector Machine(SVM) is a supervised machine learning approach that can be both used for both continuous and discrete value. In this study, it works on continuous values as a Regression model. It consists of the kernel which is used for mapping of lower dimension data to higher dimension data. A hyperplane is a decision plane that divides the set of an object into different classes, a margin which is the gap between the support vectors and Support Vectors which are the line passing through the objects. More margin gap means a good margin and less margin gap means bad margin. Since the output is a real number, the objective is to minimize the error.

SVM regression function represented by w T x.Support vectors are indicated by a red dot and "Ɛ" is the distance. Here in this study, Y - variable represents the total number of cases and the X variable represents the date, Fig. 2.

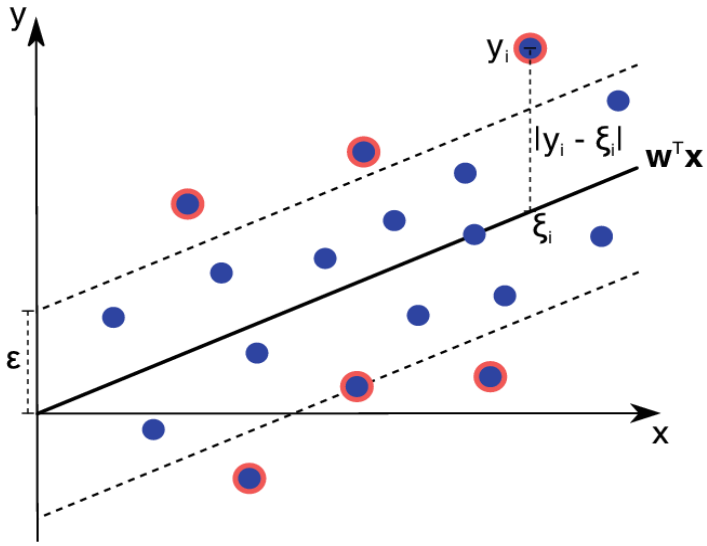


Fig.2. Support Vector Machine architecture

C. Random Forest Regression (RFR)

Random Forest Regression(RFR) model is an ensemble learning method that uses multiple decision trees, splitting nodes in each tree considering the limited number of features and using multiple machine learning algorithms together to make an accurate prediction and this technique known as bagging. Prediction is made by averaging all the predictions of the decision tree. In this study, the Y - variable represents the total number of cases and the X variable represents date is used to predict the output .Fig.3.

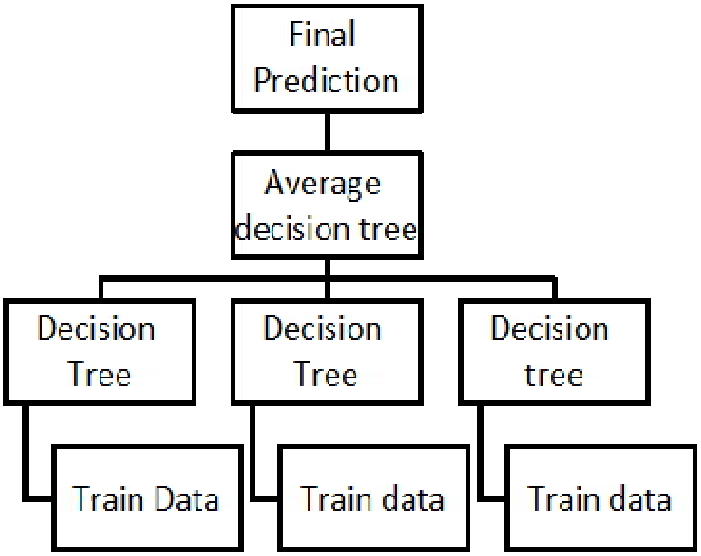


Fig. 3. Random Forest Regression architecture

D. XGBoost Regression (XGB)

XGBoost (eXtreme Gradient Boosting) model is used in decision tree-based ensemble learning algorithms that use a gradient boosting framework.XGBoost algorithm was developed as a research project at the University of Washington. It has an additional feature for doing cross-validation and computing feature importance. Boosting is one of the techniques that train models in succession and corrects the error made by the previous model. The X variable represents the date and the Y variable represents the total number of cases is used to predict the output, Fig.4.

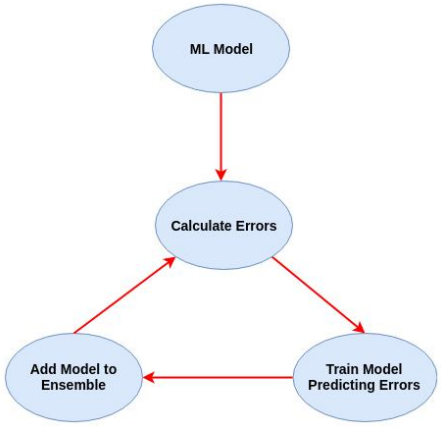


Fig. 4.XGBoost Regression architecture

III.RESULTS AND DISCUSSION

A. Data Information

Our world in Data COVID-19(OWID) dataset by the University of Oxford is being updated daily and published by the European CDC. The total cases registered are 69,60,259. The parameters mainly used to predict the outbreak are Date, Total cases.

B. Performance Metrics

The performance metrics used in this study are MAE(Mean absolute error) and RMSE(root mean square error).

1) *Mean Absolute Error (MAE):* MAE is used to measure the accuracy of a continuous variable. It measures the average magnitude of errors for a set of predictions. It is the result of the difference between two continuous variables.Eq.2.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Eq.2. Mean absolute error equation

2) *Root Mean Squared Error:* RMSE is the square root of MSE. It measures the values predicted by between the hypothetical models and the observed values. It is most useful when large errors are particularly desirable.Eq.3.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Eq.3. Root mean squared error equation

C. Best Predictor Model

Choosing the best predictor model in comparison with the LR, SVR, RFR, XGB.For performance, Root Mean Squared Error(RMSE), Mean Absolute Error (MAE), and Score parameters of LR, SVR, RFR, and XGB are taken into consideration for comparing the models. These errors are calculated out of a total of 69,60,259 cases.

For MAE parameter LR gives 7,01,352.85,SVR gives 7,36,615.06,RFR
gives 6,97,490.74 and XGB gives 6,96,677.62 errors .

For RMSE parameter LR gives 7,02,894.91,SVR gives 7,36,616.55,RFR gives 7,00,010.98 and 6,99,499.64 errors.

For Score, LR gives 0.5837,SVR gives 0. 9737,RFR gives 0.9964 and XGB gives 0.9967.

D. Performance Visualization

1) Mean Absolute Error (MAE) Analysis

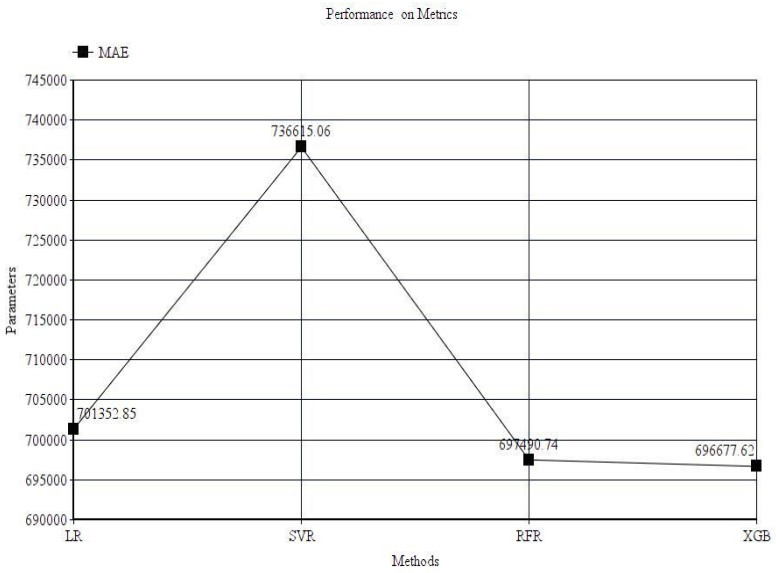


Fig. 4. Performance measurement using MAE

2) Root Mean Squared Error(RMSE) Analysis

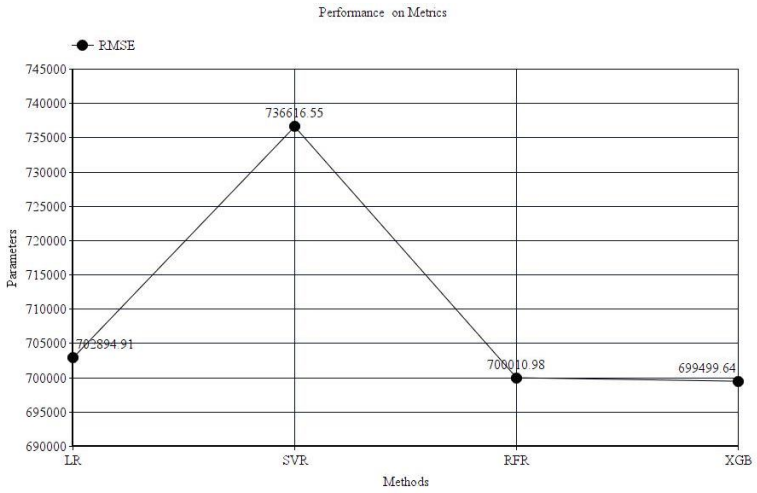


Fig. 5. Performance measurement using RMSE

3) Score Analysis

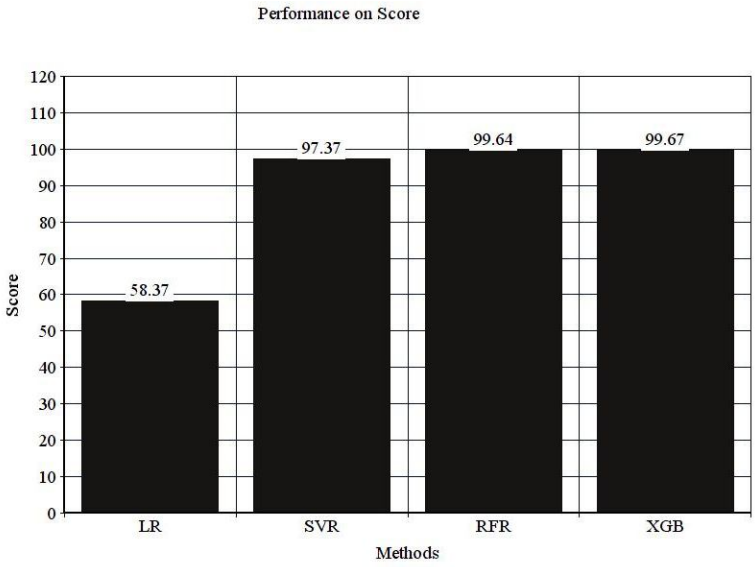


Fig. 6.Performance measurement using Score

E. Performance Tabulation

Table I
Performance tabulation of MAE,RMSE AND SCORE

Performance / Method	Linear Regression (LR)	Support Vector Regression (SVR)	Random forest Regression (RFR)	XGBoost Regression (XGB)
MAE (out of 69,60,259)	7,01,352.85	7,36,615.06	6,97,490.74	6,96,677.62
RMSE (out of 69,60,259)	7,02,894.91	7,36,616.55	7,00,010.98	6,99,499.64
SCORE(in %)	58.37	97.37	99.64	99.67

F. Calculations

Avg. Of i . MAE= Sum of MAE/total no. Methods used = 28,32,136.27 / 4 = 7,08,034.06
ii. RMSE= Sum of RMSE/total no. RMSE = 28,39,022.08 / 4 = 7,09,755.52

Total no. of cases = 69,60,259
% Error i. MAE = (Avg. Of MAE /Total no. of cases)*100 =10.17 %
ii. RMSE = (Avg. Of RMSE /Total no. of cases)*100 =10.19 %

For the best model, the least value of errors and the highest value of accuracy is considered.Analyzing the performance metrics using the MAE performance of XGB is best compare to RFR than followed by LR and SVR. The same is the outcome of using RMSE.
Using Score ,The performance of XGB is best and almost comparable to RFR than followed by SVR and LR.
Using all the 3 metrics, XGB performs the best compare to RFR. Between SVR and LR, the error is 10% (approx.) for the models used in this study, we cannot conclude the final performance using these 2 metrics. Since there is a huge difference in performance using Score of about 39%. Thus, the performance of SVR is considered better than LR.

As in LR, there is non-linearity in the predicted observation and the actual observation. Thus, the model fails to perform well for prediction and have low score compare to SVR, where it fits many predicted observation into the support vectors of actual observation.RFR model fits the non-linearity because each model in the decision tree guides the next model to focus on those particular features. Thus, fitting the non-linearity in a much better way compared to SVR.XGB model fits the non-linearity model perfectly because each decision tree's weak performance, which runs parallel are aggregated to obtain final prediction which is better than any individual prediction themselves. Thus, it minimizes the regularised objective function based on predicted and actual observation. Hence, it performs comparably better than RFR.

Overall the final result shows that the performance of XGB is better than RFR followed by SVR and LR.

IV. CONCLUSIONS

In the study, efforts were made to use machine learning algorithms with Our World in Data COVID-19 data to predict the outbreak and analyze which method gives better performance. In conclusion, the performance of XGBoost(XGB) is better than Random Forest Regression(RFR) followed by Support Vector regression (SVR) and Linear Regression(LR).The observation is made using Our world in the Data COVID-19 dataset. It has been observed that using more accurate and more data, the error can further be reduced.

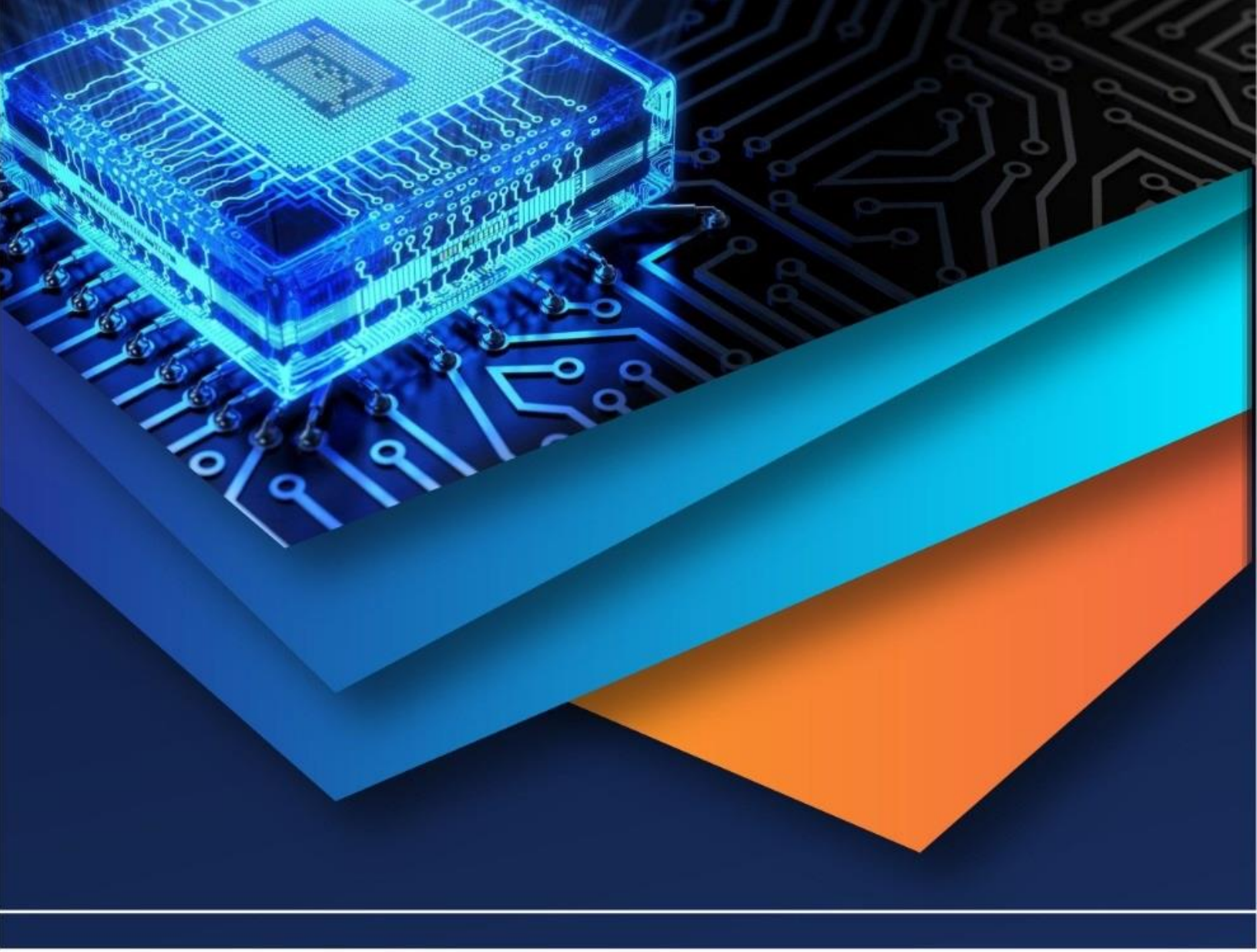
The study can be further extended in the future using different datasets or using different methods like Artificial Neural Networks(ANN), Extreme machine learning (ELM), or (Classification and Regression Trees (CART) for better performance of the model.

V. ACKNOWLEDGMENT

My sincere gratitude to DR.C.N. Subalalitha, Professor, SRM institute of technology for guidance, help, advice, and support throughout this study without her this study would not have possible.

REFERENCES

[1] Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P.B., Joe, B. and Cheng, X., 2020. Artificial intelligence and machine learning to fight COVID-19.
[2] Wynants, L., Van Calster, B., Bonten, M.M., Collins, G.S., Debray, T.P., De Vos, M., Haller, M.C., Heinze, G., Moons, K.G., Riley, R.D. and Schuit, E., 2020. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. bmj, 369.
[3] cYan, L., Zhang, H.T., Xiao, Y., Wang, M., Sun, C., Liang, J., Li, S., Zhang, M., Guo, Y., Xiao, Y. and Tang, X., 2020. Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. MedRxiv.
[4] Yang, Z., Zeng, Z., Wang, K., Wong, S.S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z. and Liang, J., 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. Journal of Thoracic Disease, 12(3), p.165.
[5] Long, B., Bian, J., Dong, A. and Chang, Y., 2012, October. Enhancing product search by best-selling prediction in e-commerce. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 2479-2482).
[6] Nguyen, D., Smith, N.A. and Rose, C., 2011, June. Author age prediction from text using linear regression. In Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities (pp. 115-123).
[7] Han, Z., Liu, Y., Zhao, J. and Wang, W., 2012. Real time prediction for converter gas tank levels based on multi-output least square support vector regressor. Control Engineering Practice, 20(12), pp.1400-1409.
[8] Asim, K.M., Idris, A., Iqbal, T. and Martínez-Álvarez, F., 2018. Earthquake prediction model using support vector regressor and hybrid neural networks. PloS one, 13(7), p.e0199004.
[9] Kane, M.J., Price, N., Scotch, M. and Rabinowitz, P., 2014. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC bioinformatics, 15(1), p.276.
[10] Jog, A., Carass, A., Roy, S., Pham, D.L. and Prince, J.L., 2017. Random forest regression for magnetic resonance image synthesis. Medical image analysis, 35, pp.475-488.
[11] Ogunleye, A.A. and Qing-Guo, W., 2019. XGBoost model for chronic kidney disease diagnosis. IEEE/ACM transactions on computational biology and bioinformatics.
[12] Ardabili, Sina & Mosavi, Amir & Ghamisi, Pedram & Ferdinand, Filip & Varkonyi-Koczy, Annamaria & Reuter, Uwe & Rabczuk, Timon & Atkinson, Peter. (2020). COVID-19 Outbreak Prediction with Machine Learning. 10.20944/preprints202004.0311.v1.
[13] Li, L., Situ, R., Gao, J., Yang, Z. and Liu, W., 2017, October. A hybrid model combining convolutional neural network with xgboost for predicting social media popularity. In Proceedings of the 25th ACM international conference on Multimedia (pp. 1912-1917).
[14] Cheng, F.Y., Joshi, H., Tandon, P., Freeman, R., Reich, D.L., Mazumdar, M., Kohli-Seth, R., Levin, M., Timsina, P. and Kia, A., 2020. Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients. Journal of Clinical Medicine, 9(6), p.1668.
[15] Kukar, M., Gunčar, G., Vovko, T., Podnar, S., Černelč, P., Brvar, M., Zalaznik, M., Notar, M., Moškon, S. and Notar, M., 2020. COVID-19 diagnosis by routine blood tests using machine learning. arXiv preprint arXiv:2006.03476.
[16] Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q. and Cao, K., 2020. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology, p.200905.
[17] Narin, A., Kaya, C. and Pamuk, Z., 2020. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. arXiv preprint arXiv:2003.10849.
[18] Pandey, G., Chaudhary, P., Gupta, R. and Pal, S., 2020. SEIR and Regression Model based COVID-19 outbreak predictions in India. arXiv preprint arXiv:2004.00958.
[19] Wang, L. and Wong, A., 2020. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. arXiv preprint arXiv:2003.09871.
[20] Wang, L., Han, D. W., Li, K., Yang, X., Yin, X. L., Qiu, J., ... & Ma, Z. Y. (2020). A Universal Model for Prediction of COVID-19 Pandemic Based on Machine Learning.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY